



ELSEVIER

Chemometrics and Intelligent Laboratory Systems 28 (1995) 149–164

Chemometrics and  
intelligent  
laboratory systems

## Spectral curve fitting of infrared spectra obtained from semi-crystalline polyester yarns

A.P. de Weijer <sup>a,1</sup>, L. Buydens <sup>a,\*</sup>, G. Kateman <sup>a</sup>, H.M. Heuvel <sup>b</sup>

<sup>a</sup> Department of Analytical Chemistry, Catholic University of Nijmegen, Toernooiveld 1, NL-6525 ED Nijmegen, The Netherlands

<sup>b</sup> Akzo Nobel Central Research, P.O. Box 9300, NL-6800 SB Arnhem, The Netherlands

Received 8 September 1994; accepted 7 December 1994

### Abstract

Infrared spectra of polymer yarns exhibit a large number of strongly overlapping peaks. One of the methods to study the physical yarn structure quantitatively is to resolve the spectra into separate bell-shaped peaks. Due to the overlapping peak patterns, curve fitting with steepest descent like procedures depend on accurate initial estimates of the peak parameters. We developed a non-local curve fitting procedure, based on a genetic algorithm, which proved to be less sensitive to initial estimates. The combination of a genetic algorithm with a nonlinear estimator, based on Marquardt's method, gave the best curve fitting results.

### 1. Introduction

Since the introduction of fast Fourier transform spectrometers, the quality of spectra has improved considerably. In many applications of vibrational spectroscopy the analyst is interested not only in the qualitative chemical nature of the investigated sample, but also in the quantitative determination of basic aspects of polymeric structure: state of order, configurational and conformational regularity, molecular orientation, etc. Sets of high-quality spectra open possibilities to obtain direct and accurate information on the physical or chemical state of sub-

stances from curve resolution of spectra. However, proper and unambiguous fitting, necessary for obtaining accurate quantitative measures, is not simple. The steps to resolve a peak pattern in its individual components involve two major activities. First, estimates of the number of peaks and their corresponding peak positions, widths at half height and intensities have to be determined to build the mathematical model for the subsequent next step, the parameter optimization. Since the mathematical model is non-linear and since it cannot be solved analytically, a search procedure is required. This second step can either be served by 'traditional' steepest descent like procedures or by non-local search procedures such as genetic algorithms (GAs). To illustrate the present-day possibilities in this respect, the development of a peak-fit model and the curve fitting of infrared spectra of poly(ethylene terephthalate) (PET) yarns is described.

\* Corresponding author.

<sup>1</sup> Permanent address: Akzo Nobel Central Research Arnhem, P.O. Box 9300, 6800 SB Arnhem, The Netherlands.

## 2. Error sources related to spectral curve fitting

Curve fitting can be a long and complicated task if the underlying mathematical model is not known, or fairly accurate initial estimates of the parameters in the model cannot be obtained. Moreover, a close fit to the experimental profile does not always lead to physical relevant peak parameters if the problem is overdetermined; i.e. many sets of parameters may give equal close fits to the experimental spectrum. A quantitative study on the influences of the degree of overlap, the number of unresolved bands on the fitting results of theoretical and experimental infrared spectra was given by Vandeginste and De Galan [1].

The main error sources of curve fitting can be classified into three categories [2]: (1) The exact number of peaks or peak shape is not known, leading to poor convergence or to ambiguous solutions. (2) Uncertainty about the baseline position. (3) Initial estimates of the parameters in the model are insufficiently accurate; this may cause the problem that the fitting procedure ends in a local optimum.

Error sources 1 and 2 originate from the experimental data, including the way the latter are collected and prepared for the subsequent optimization

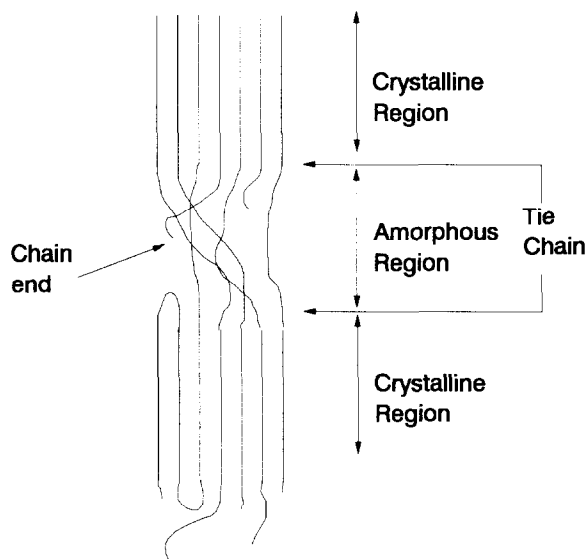


Fig. 1. Two-phase model for a PET fiber.

method used for curve fitting. In contrast, error source 3 originates from the optimization method.

Traditionally, local searching optimization methods are most widely applied, e.g. the Gauss–Newton minimization method for nonlinear models. In con-

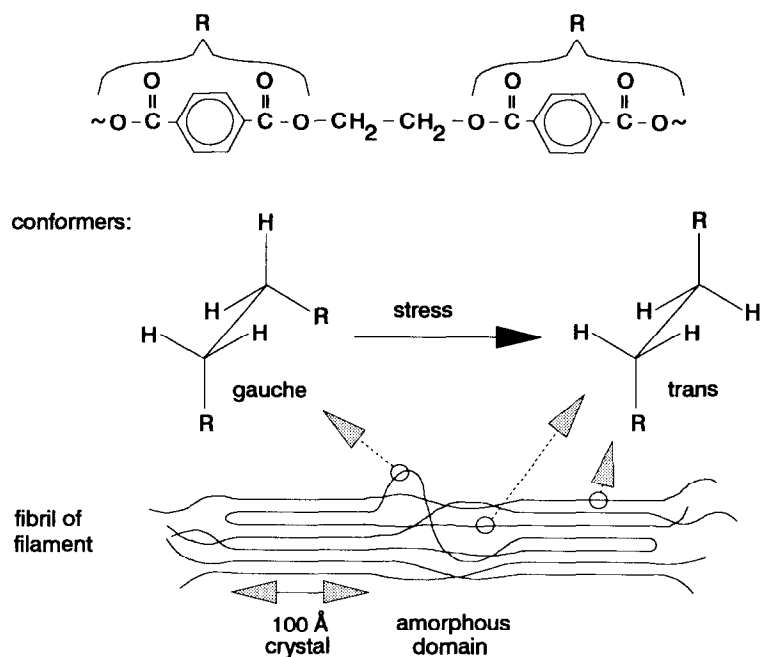


Fig. 2. Molecular and physical structure of PET.

trast, globally searching optimization methods are more robust, i.e. less sensitive to initial estimates. Admittedly, the latter methods are computationally more intensive as well, but with the advent of increasingly faster and cheaper computers, computation times within practical limits become feasible.

### 3. Field of application; semi-crystalline polymers

The physical structure of semi-crystalline polymers can be described adequately by a two-phase model, consisting of alternating amorphous and crystalline regions. Using a combination of measuring techniques such as density, X-ray diffraction, pulse propagation, birefringence, and thermal analysis, a good impression of the parameters involved in such a model can be obtained [3]. These parameters proved to be quite useful in understanding the relations between process conditions, physical yarn structure and (thermo)mechanical properties [4,5]. A problem is to obtain a more detailed picture of the amorphous regions. Whereas the structure of the crystals can be studied in detail, using X-ray diffraction techniques, the information about the amorphous regions is limited through rough overall quantities such as the average orientation of the molecules or the average size of the amorphous regions. In Fig. 1 the usual model for an amorphous region as encapsulated between two crystals is depicted. As a first approximation, the crystals can be considered as stiff blocks. Therefore, it is clear that the molecular arrangement in the amorphous regions determines the mechanical behaviour of the yarn to a large extent. A quantitative study on these relations has been described elsewhere [4]. From the literature, evidence has emerged that infrared spectroscopy is able to provide more detailed information on structural features of the amorphous regions [6–8] as shown in the following two examples.

The ethylene groups in the amorphous domains of semi-crystalline PET occur in two conformations, namely the *gauche* and the *trans* conformation. As can be concluded from Fig. 2, molecular chains containing a lot of *gauche* will be coiled strongly, whereas *trans* conformers in series give rise to extended chains. Clearly, crystals only contain *trans*. In the infrared spectrum, the amount of *gauche* confor-

mation can be quantified by the area of the CH<sub>2</sub> rocking vibration band at about 1045 cm<sup>-1</sup> and 899 cm<sup>-1</sup>. Furthermore, the crystalline and amorphous *trans* conformers show characteristic absorptions at about 970 cm<sup>-1</sup> as a result of the C–O stretching mode.

Polarized infrared spectroscopy may give information on the (local) orientational order parameters of polymer fibers. From the dichroism of an infrared band, the orientation of that part of the molecule which takes part in the vibration mode involved can be obtained. The dichroic ratios are calculated as

$$D = \frac{A_{\parallel}}{A_{\perp}}$$

with  $A_{\parallel}$  and  $A_{\perp}$  indicating the integrated area of the absorbance peak for the parallel and perpendicular polarized spectra, respectively. From  $D$ , the Herman's orientation function can be calculated:

$$f = \frac{3\langle \cos^2\theta \rangle - 1}{2} = \frac{(D - 1)(D_0 + 2)}{(D + 2)(D_0 - 1)}$$

where  $D_0$  depends on the angle  $\alpha$  between the transition moment and the molecular axis.  $\theta$  represents the angle between the polymer-chain axis and the fiber axis

$$D_0 = 2 \cot^2 \alpha$$

If  $\alpha = 54.7^\circ$  (magic angle),  $f$  is undefined. Positive orientation factors correspond with  $D$  values  $> 1$  when  $\alpha < 54.7^\circ$  ( $\pi$  dichroic character).  $D$  values between 1 and 0 correspond with  $\alpha > 54.7^\circ$  ( $\sigma$  dichroic character). Most of the bands in the region 700–1100 cm<sup>-1</sup> have  $\pi$  dichroic character since the chain and the transition moments are more or less parallel oriented to the fiber axis. Three bands with  $\sigma$  dichroism have been reported for PET at 1043, 875 and 730 cm<sup>-1</sup> [6,19].

Since the quantitative characterization of the physical state and orientational order is directly related to areas of bands, it is expected that the accuracy of characterization is directly related to the accuracy of estimation of the peak areas. In general, IR spectra contain a large number of strongly overlapping bands, which makes direct estimation of individual peak areas impossible. Therefore an accurate and robust curve fitting algorithm is needed.

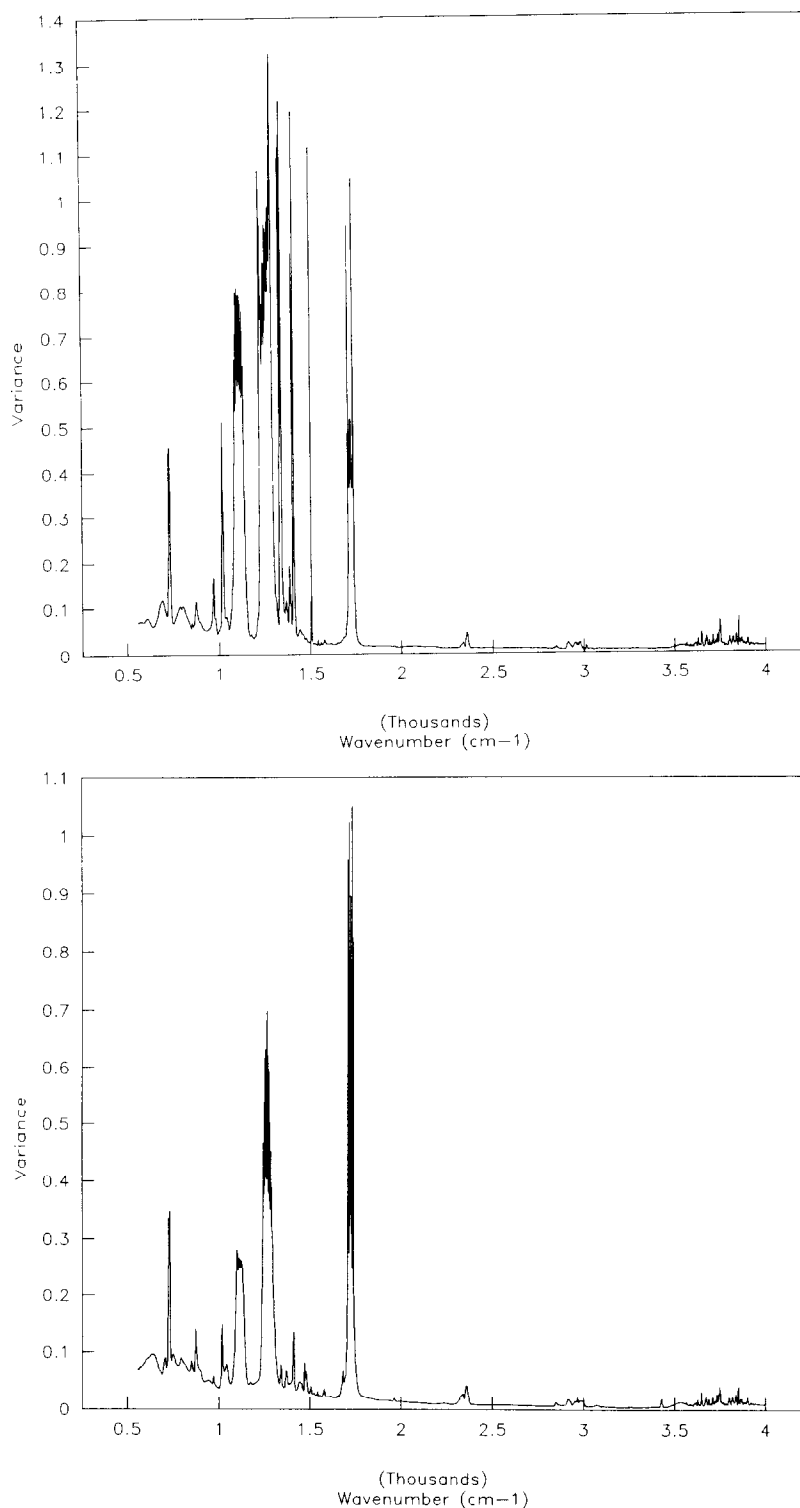


Fig. 3. Variance of 94 parallel (upper) and perpendicular (lower) polarized infrared spectra of PET. Variances higher than 0.5 originate from parts in the spectrum with too low transmission values; so in these regions variances are not responses of physical structure.

### 3.1. Samples

A subset of 94 representative PET yarns was selected from 295 yarns, originating from a large experiment in which process conditions such as chip viscosity, winding speed, draw ratio, drawing temperature, and relaxation ratio had been varied. The physical structure of all these 94 yarns was characterized by five structure parameters [4]. These parameters were calculated from wide-angle X-ray diffraction patterns, density measurements, pulse propagation velocities and relative viscosity measurements and comprise: (1) the volume fraction of crystalline material ( $V_c$ ); (2) the yarn viscosity as a measure of the length of the molecules ( $\eta_{rel}$ ); (3) an overall amorphous orientation factor ( $F_{as}$ ); (4) a measure for the contour-length distribution factor of the tie chains between successive crystals ( $F_{ad}$ ); (5) the average volume of the individual amorphous regions ( $S_a$ ).

The samples were prepared for the Fourier transform infrared (FTIR) experiment by winding the yarn on a metal frame so that a smooth layer of approximately ten parallel monofilaments is formed. Specimens were examined using a Perkin Elmer 1800 Fourier transform infrared spectrometer with a nitrogen cooled MCT detector. From each sample parallel and perpendicular polarized spectra were recorded using a Spectra Tech Microscope accessory for PE 1800 with a spectral resolution of  $4\text{ cm}^{-1}$ . In Fig. 3 the variance of parallel and perpendicular spectra over all samples is shown for the region  $4000\text{--}600\text{ cm}^{-1}$ . It shows that the area of highest variance (originating from different physical structures) occurs in the region from  $1100\text{--}700\text{ cm}^{-1}$ . This is in agreement with earlier conclusions [6]. Therefore, fitting of the infrared spectra was confined to this region. The variances were calculated in the following way:

$$s^2(\nu) = \frac{\sum_{i=1}^{94} [I_i(\nu) - \bar{I}(\nu)]^2}{93}$$

## 4. Development of the mathematical model

As a mathematical peak model, a sum of several Pearson VII lines was used [9]. On a theoretical ba-

sis IR absorption bands have Lorentzian profiles. However, a number of factors are known that result in a Gaussian broadening of the Lorentzian profile. Some of them have a chemical origin, others occur from the measurement circumstances. The Pearson VII function provides the possibility of using a wide variety of lineshapes from Gaussian to Lorentzian and beyond [9] and was therefore adopted to build the model.

$$I(\nu) = \sum_{i=0}^n \frac{I_{i,0}}{\{1 + 4[(\nu - \nu_{i,0})/H_i]^2(2^{2-Z_i} - 1)\}^{1/(2-Z_i)}} + B(\nu)$$

where  $I(\nu)$  = estimated spectrum;  $I_{i,0}$  = intensity in the center of the peak  $i$ ;  $\nu_{i,0}$  = peak position of peak  $i$ ;  $H_i$  = half-width of peak  $i$ ;  $Z_i$  = tailing factor of peak  $i$ ;  $n$  = number of peaks;  $B(\nu)$  = baseline.

### 4.1. Use of domain knowledge

The infrared spectrum of PET has been well studied by many workers [10–12]. There are some differences in the vibrational assignments regarding certain IR bands in the polymer. It can be expected that peak positions of absorption bands corresponding to skeletal backbone vibrations would be affected by internal stress. These frequency shifts may not always be detectable due to overlapping of other bands. However, the use of domain knowledge is useful in determining approximate peak positions for a number of bands.

### 4.2. Peak detection using artificial neural networks

In an earlier publication, we presented a method for peak detection, based on artificial neural networks (ANN) [13]. This ANN is capable of detecting peaks in a spectrum. ANNs are now frequently subjected to calibrate nonlinear relations and to pattern recognition with great success [13–16]. The ANN discussed here has been trained to estimate the number of bands, peak position(s) and bandwidth(s) of experimental spectra. A series of synthetic patterns of peaks varying from Lorentzian to Gaussian shapes with different degrees of overlap is used to train the ANN. The ANN, trained to recognize  $\nu_0$  in overlapping peak patterns, uses the first and second deriva-

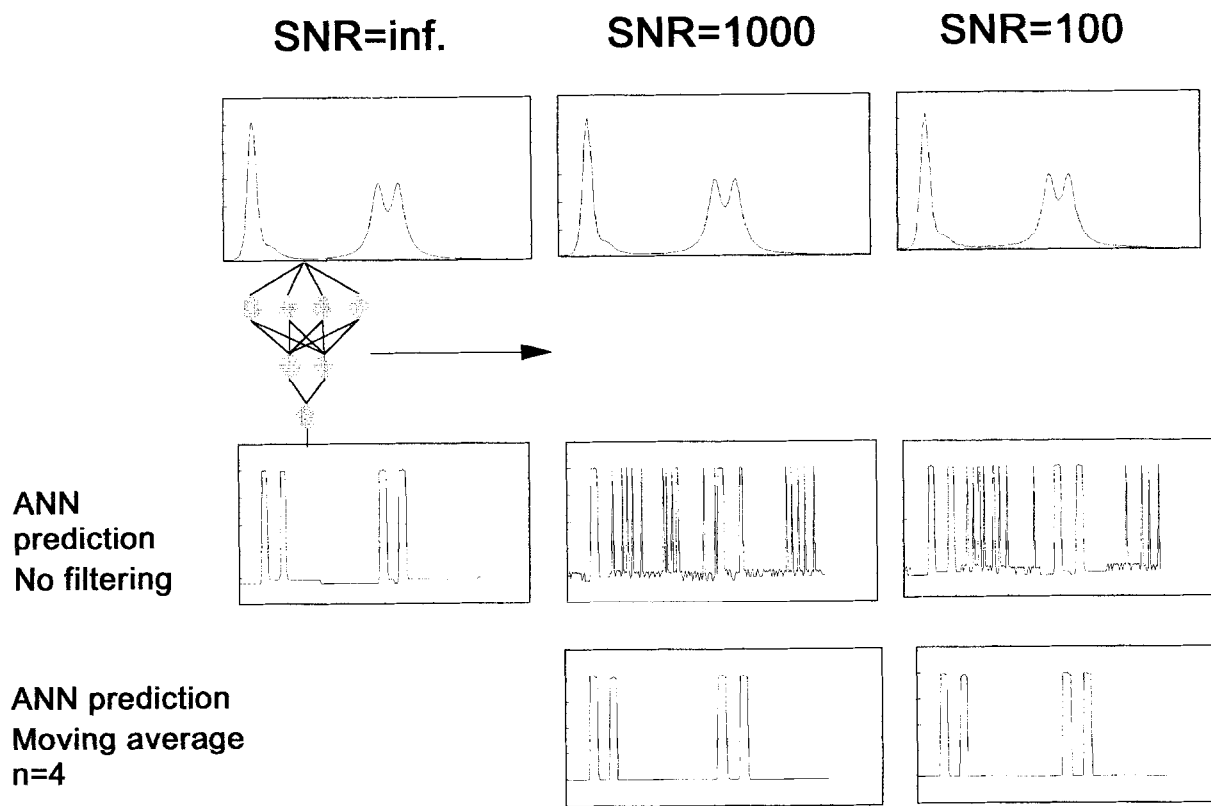


Fig. 4. A trained ANN is used to detect peaks by scanning the profile. Since this ANN depends on numerical derivatives, it is sensitive to noise. We merely applied a moving average filter to the original profile.

tives of  $I(\nu)$  to predict the presence or absence of a peak in  $\nu$ . The relation between these parameters and the presence or absence of a peak was learned by an ANN and not derived from mathematical formulations. Although the training set was set up using synthetic spectra without noise, the ANN was capable of detecting peaks in noisy synthetic peak patterns (Fig. 4) and in experimental peak patterns with different degree of overlap, provided that some noise was removed; to that end, a simple filter sufficed. More information is given in Ref. [13].

As an illustration, an experimental IR spectrum of a parallel polarized poly(ethylene terephthalate) is shown in the upper graph in Fig. 5. The estimation of some peak parameters is accomplished by scanning of the original profile employing a trained artificial neural network. The signal of the output neu-

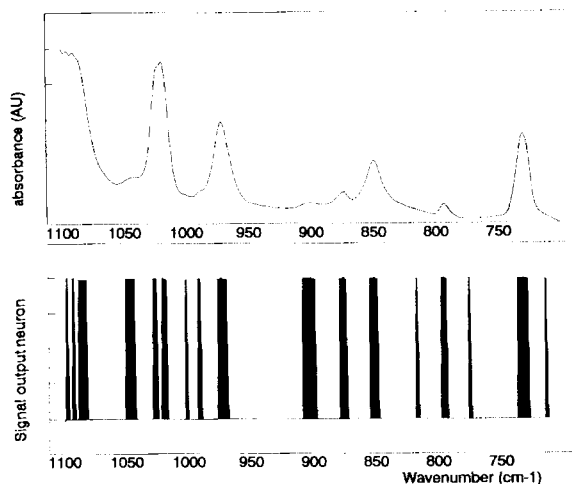


Fig. 5. IR spectrum of poly(ethylene terephthalate) yarn (top). Signal of the ANN for peak detection (bottom).

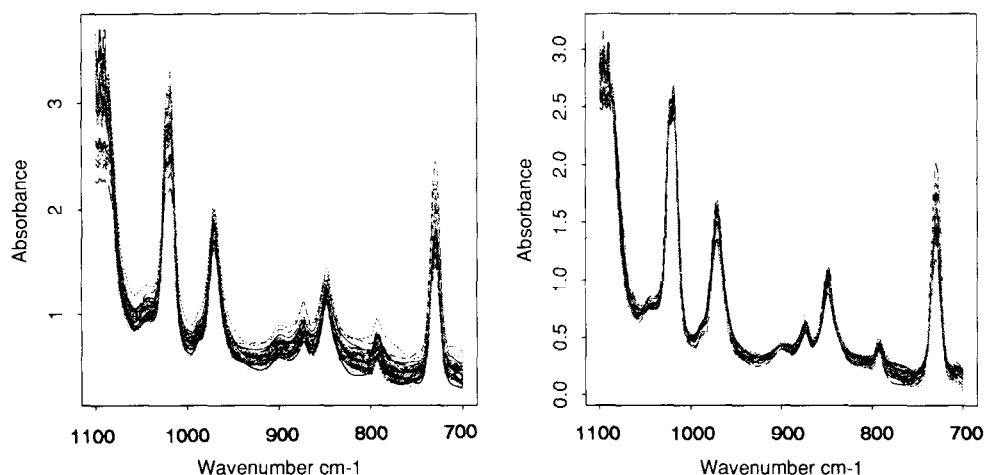


Fig. 6. Superimposing the 94 parallel polarized spectra before (left) and after (right) MSC.

ron is shown in the lower graph in Fig. 6. It shows that the output neuron triggers when the center of the scanning window is positioned between the half-widths of an (underlying) peak. From the signal of the output neuron, estimates of peak positions and half-widths can be derived.

#### 4.3. Baseline compensation with multiplicative scattering correction (MSC)

MSC has proved to be a useful correction method for absorbance shift in NIR reflectance spectroscopy [19]. Although MSC is originally applied to compen-

sate for scattered straylight, it is also useful for compensation of baselines of transmittance IR as long as the variations originating from variations in physical structure are smaller than the variations due to baseline shifts. First, linear transformations are usually applied. The most common transformation is to transform transmission infrared spectra ( $T(\nu)$ ) into absorbance spectra ( $I(\nu)$ ) in the following way:

$$I(\nu) = -\log(T(\nu))$$

To eliminate the differences in baseline levels, MSC uses linear regression on spectral variables ver-

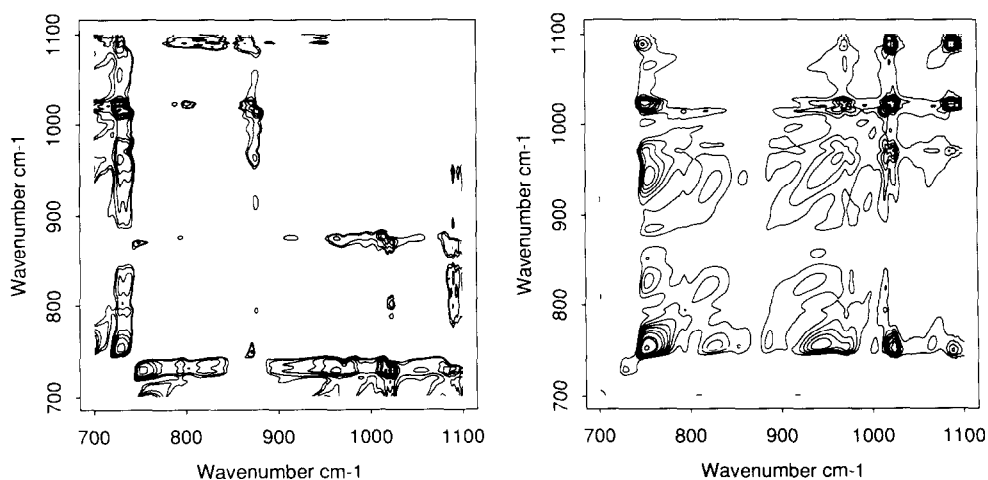


Fig. 7. Two-dimensional covariance plot of 94 dichroic spectra. The negative covariances are shown in the left plot, positive covariances are shown in the right plot.

sus average spectrum. If  $I(\nu)$  is the (linearly transformed) spectrum of a sample and  $\bar{I}(\nu)$  is the average (linearly transformed) spectrum, the parameters  $\alpha$  and  $\beta$  can be estimated by least-squares fit.

$$I(\nu) = \bar{I}(\nu) \alpha + \beta + \varepsilon$$

$I(\nu)_{\text{corr}}$  is equal to

$$I(\nu)_{\text{corr}} = \frac{I(\nu) - \beta}{\alpha}$$

$\varepsilon$  contains the residual noise and structure information. MSC is applied to compensate for local optical pathway differences (i.e. local differences in yarn diameter or internal reflections). These differences contribute to a transmission shift. The problem of fitting different baselines to each spectrum is reduced to the problem of fitting one baseline to all spectra. MSC correction for parallel polarized spectra of PET is shown in Fig. 6.

#### 4.4. Peak detection on series using two-dimensional covariance calculations

Recently, a general two-dimensional correlation method for infrared and Raman spectroscopy was introduced by Noda [17]. This method comprises the exertion of an external perturbation on a system while being monitored by an external electromagnetic probe. From the obtained time-resolved spectra,

two-dimensional spectra can be calculated. It contains information about time-dependent interactions of the physical state of the system. A similar time-independent approach is proposed by Barton et al. [18]. Here, a two-dimensional plot of correlation coefficients is calculated from the deviations with respect to the mean values of IR intensities at different wavenumbers. This type of correlation-coefficient plot, based on a series of samples, always has constant value of unity at the diagonal position. For peak detection purposes it is more convenient to retain spectral information. Multiplication of the correlation matrix by the deviation matrix results in a covariance matrix. The diagonal represents the variance spectrum. The intensity of the two-dimensional covariance spectrum becomes

$$C(\nu_1, \nu_2) = \sum_{i=0}^n \frac{[I(\nu_1) - \bar{I}(\nu_1)][I(\nu_2) - \bar{I}(\nu_2)]}{n-1}$$

When the number of spectra ( $n$ ) is large and the variation in peak position is relatively small, two-dimensional covariance calculations is useful in the development of a suitable model for curve fitting. Since peaks with zero variance do not contribute to the intensity of the covariance signals, the two-dimensional covariance matrix may contain highly resolved peak patterns of structure dependent peaks. Moreover, information concerning correlated peaks can directly be obtained from the plot. To illustrate

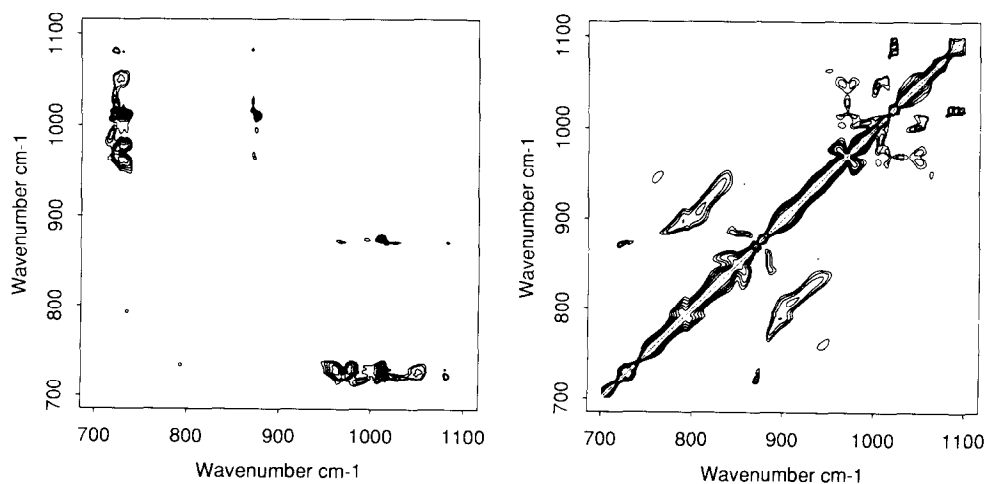


Fig. 8. Same as Fig. 7, only for correlation coefficients. Correlation coefficients  $> 0.75$  and  $< -0.75$  are displayed.



this, the two-dimensional covariance and correlation matrix for dichroic spectra, in which  $I_d(\nu) = I_{\parallel}(\nu)/I_{\perp}(\nu)$ , is shown in Figs. 7 and 8, respectively. Only the correlation coefficients  $< -0.75$  (Fig. 8a) and  $> 0.75$  (Fig. 8b) are plotted. For isotropic bands the  $I_d(\nu) = 1$ ; there is no preferable orientation direction.  $I_d(\nu)$  decreases with orientation of the chain segment for bands with  $\sigma$  character and increases for  $\pi$  bands. Figs. 7a and 7b both display a strong negative coupling between  $875\text{ cm}^{-1}$ ,  $730\text{ cm}^{-1}$   $\sigma$  bands and other  $\pi$  bands. As indicated by the + covariance plot (Fig. 7b) as well as by the + correlation plot the absorption band near  $1020\text{ cm}^{-1}$  is composed of two bands, at  $1018\text{ cm}^{-1}$  and at  $1023\text{ cm}^{-1}$  both positively coupled to the  $972\text{ cm}^{-1}$  band. In Fig. 9 an intersection of Fig. 7 at  $972\text{ cm}^{-1}$  shows this more clearly. The composition of the  $1020\text{ cm}^{-1}$  band is confirmed by Yang [11] who assigned the peaks to different conformers of the ethylene glycol. They also reported on the presence of a broad band near  $1022$

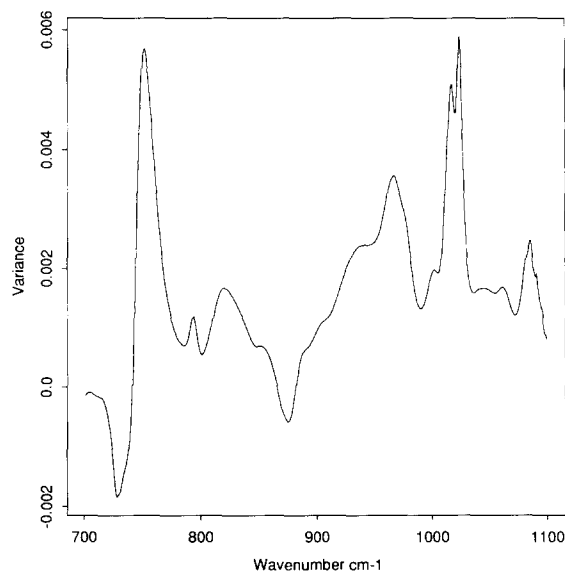


Fig. 9. Intersection at  $972\text{ cm}^{-1}$  of the two-dimensional covariance plot as shown in Fig. 7. The band at  $1020\text{ cm}^{-1}$  splits up.

Table 1  
Initial estimates derived from three pre-hybridization procedures

Band number	Detection			Estimates				
	A	B	C	$\nu_0$	$I_{0\parallel}$	$I_{0\perp}$	$H_0$	Z
1	*	*	*	1090.5	3.000	3.000	31.0	1.5
2	*	*	*	1046.8	0.458	0.248	28.2	1.5
3	*	*	*	1023.8	1.116	0.049	9.9	1.5
4	*			1018.5	0.435	0.205	46.2	1.98
5	*	*	*	1017.6	1.139	0.397	8.3	1.5
6	*		*	1002.2	0.050	0.030	15.0	1.98
7	*	*	*	989.4	0.057	0.039	7.3	1.5
8	*		*	972.0	0.192	0.161	36.2	1.98
9	*	*	*	971.5	1.058	0.168	14.3	1.5
10		*		945.0	0.127	0.066	38.0	1.98
11	*	*	*	898.4	0.242	0.119	21.5	1.98
12			*	878.0	0.178	0.142	22.9	1.98
13		*		875.0	0.106	0.406	13.7	1.98
14	*			873.1	0.078	0.572	6.0	1.98
15	*	*	*	849.0	0.420	0.239	11.4	1.5
16			*	851.1	0.427	0.320	34.7	1.98
17		*	*	813.8	0.176	0.099	38.0	1.98
18		*	*	793.8	0.193	0.100	9.3	1.5
19		*		789.3	0.049	0.013	7.6	1.98
20			*	783.0	0.036	0.046	21.5	1.98
21		*		765.5	0.134	0.023	55.0	1.98
22	*	*	*	730.0	1.800	2.200	20.0	1.5

A = Domain knowledge; B = peak detection with artificial neural networks; C = two-dimensional covariance spectrum.

$\text{cm}^{-1}$  originating from the same conformers in the amorphous regions of PET yarns. Although this peak could not be detected with our method using this specific data base, we added the band to our first estimate of the mathematical model. The area around the *trans* C–O–C stretching vibration ( $972 \text{ cm}^{-1}$ ) is rather complex (Fig. 8b). The + covariance plot (Fig. 7b) suggests that the band consists of a sharp ‘crystalline *trans*’ peak and an asymmetric broad ‘amorphous *trans*’ band. Since the vibration is a backbone vibration, internal stresses may cause an asymmetric peak shift towards lower wavenumbers. Earlier studies have shown that the  $972 \text{ cm}^{-1}$  band is extremely sensitive to strain [6,19].

Some information about orientational order and *trans/gauche* peaks could be derived directly from the plots. Although there are some recent developments in this field [20], interpretation in terms of pure variables such as crystallinity and orientation distribution remains difficult. The + covariance and the – covariance spectra of dichroic as well as isotropic spectra could easily be used for peak detection purposes. Using the same procedure as described in Section 4.2, all two-dimensional plots were scanned by the neural network by taking all intersections. The output signal of the ANN is used to reconstruct the two-dimensional plots from which peaks and corresponding peak positions were taken by picking every block of ANN responses. Spikes were excluded.

In the previous sections we considered three methods which can be useful in finding initial estimates for the mathematical model, i.e. the pre-hybridization phase. These estimates can either serve as a starting point for a local search procedure or can be the center of the search volume for a non-local search procedure. In Table 1, initial estimates for the fitting of parallel and perpendicular PET infrared spectra are listed. The table also shows the method with which each specific band had been detected. In many cases the methods are complementary. Some peaks have only been detected with one of the methods. Although it remains uncertain if these peaks are artefacts resulting from the computational methods that are used, we have some indications that the detected peaks originate from vibrational modes. Each vibrational mode with contributions from the amorphous and the crystalline regions gives rise to a double peak in the infrared spectrum. A broad peak at  $851 \text{ cm}^{-1}$

Table 2

Deviations from initial estimates of Table 1; these deviations form the search volume of the genetic algorithm

Peak positions ( $\nu_{0,i}$ )	$d_i = 4 \text{ cm}^{-1}$
Intensities ( $I_{0\parallel,i}$ and $I_{0\perp,i}$ )	$d_i = \frac{1}{2} I_{0\parallel,i}$ or $\frac{1}{2} I_{0\perp,i}$
Half widths ( $H_0$ )	$d_i = \frac{1}{2} H_0$
Tailing factor ( $Z$ )	if $Z = 1.5$ then $d_i = 0.5$ if $Z = 1.98$ then $d_i = 0$

and a small peak at  $849 \text{ cm}^{-1}$  have been found by two-dimensional peak detection, possibly originating from *trans* conformers in crystalline and amorphous regions. The triple peak pattern at  $875 \text{ cm}^{-1}$  is a benzene OOP vibration. Crystalline and amorphous IR-patterns for this band have been reported earlier [6]. In NMR studies associates of phenylene rings in the amorphous regions of PET have been detected [22]. But since there is no hard evidence that associates have been found in more extensive IR studies, it is not certain that this extra peak is assigned to benzene associates.

Genetic algorithms require a search volume instead of an initial setting. For the  $i$ th parameter, the value range can be defined in terms of a center value  $c_i$  (see  $\nu_0$ ,  $I_{0\parallel}$ ,  $I_{0\perp}$ ,  $H_0$ ,  $Z$  in Table 1) and a deviation  $d_i$ ; range  $R_i = [c_i - d_i, c_i + d_i]$ . The deviations for each of the Pearson VII parameters are given in Table 2.

## 5. Fitting the mathematical model to the experimental data

All search techniques have in common that they basically aim to improve an initial estimate of the true solution. Local search techniques traditionally applied to tackle curve fitting problems are relatively fast in comparison to search techniques that employ non-local search heuristics, e.g., statistical ‘cooling’ algorithms (simulated annealing) or genetic algorithms. On the other hand, the latter techniques are less sensitive to local optimal points. For both local and global searches a fitness or error function is needed. The root mean square error (RMSE) between experimental and fitted spectrum is used as an evaluation criterium:

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^n [I_j(\nu) - \hat{I}_j(\nu)]^2}{n-1}}$$

### 5.1. Gauss–Newton optimization methods

A steepest descent based curve fitting module with Gauss–Newton optimization method is successfully applied to a series of curve fitting problems in our laboratory. For brevity, we will refer to ‘GN’ whenever this method is meant. Curve fitting of both X-ray diffractometer scans of poly(ethylene terephthalate) (PET) and nylon 6 yarns, and infrared spectra of PET yarns as well, was based on GN for optimizing peak parameters with a least-squares criterion extended with a Marquardt technique for forcing convergence and a technique for handling constraints [21].

From X-ray diffractometer scans physically relevant parameters such as lattice parameters and apparent crystallite sizes of a sample can directly be calculated from peak positions and half-widths of the resolved bands. To obtain sufficiently accurate estimates of the various spectral parameters, a nonlinear curve fitting routine is necessary. For relatively simple diffractometer scans, the GN method works without problems. It operates fast and is not very sensitive to a priori estimates. As stated before, in complex peak patterns, initial estimates of the parameters in the model need to be accurate in order to prevent the fitting procedure from ending in a local optimum when using GN optimization. Moreover, we noticed that more time has to be spent on the development of a generally applicable model as the complexity of the spectra increases. The development of a 23-line model on infrared spectra for PET yarns showed that a large number of constraints on band parameters and coupling of parallel and perpendicular spectra had to be incorporated into the computer model for a correct convergence of the GN optimization [6]. We have shown earlier that in fitting complex spectra, in which many bands display large overlaps, the steepest descent approach can fail by lack of accuracy of the initial estimates of the fit parameters, resulting in sub-optimal solutions. Genetic algorithms are less sensitive to local optima in our experiments in which we successfully revealed underlying peaks in highly overlapping peak patterns of X-ray diffraction patterns of poly(ethylene naphthalate) yarns [13].

### 5.2. Curve fitting based on genetic algorithms

It is undesirable to force the optimization method to a solution by imposing constraints, if the problem

is actually not over-dimensionalized, i.e., if there exists a unique combination of parameters that leads to a residual fit in the order of magnitude of the noise. Such a decision is unavoidable when using steepest descent methods in high-dimensional problems. There is therefore a need for fitting procedures which are less sensitive to local optimal points.

Genetic algorithms comprise a powerful and increasingly adhered search methodology which embraces principles of Darwinian evolution. They are especially suitable for complex, large-scale optimization; for more details about their concepts and operation, the reader is referred to Refs. [23–28].

Genetic algorithms are in general insensitive to ending up in local optima, i.e. they tend to approach the globally optimal solution irrespective of diverse starting conditions. We therefore say that genetic algorithms are robust, i.e. feature a good search accuracy. By contrast, the search precision of genetic algorithms is poor: upon replicating runs, there is a considerable spread in the end solutions, despite the good search accuracy; the average end solution is a reliable estimate of the global solution in general. Importantly, opposite properties often apply to traditional local optimization techniques, such as the afore-mentioned steepest descent method: a good search precision and a poor search accuracy, as pointed out above. Therefore, better overall performance can be attained in a sequential combination, where the genetic algorithm generates a ‘best guess’ that serves as a starting point for subsequent improvement (refinement) by the local technique. Such (and other) hybrid techniques are important in that the constituent methods mutually supplement each other for enhanced overall performance.

### 5.3. Software and hardware

The program CFIT (CurveFITter) uses a genetic algorithm to search the global minimum on the error landscape in the space spanned by the model parameters. The routines comprising CFIT are split into two parts: a domain-dependent part (concerning the problem representation and the error function) and the domain-independent part (concerning the genetic search heuristics). By domain-independent routines are meant routines that can be used for other domains as well; for our application these were ob-

tained from the software library GATES [27]. The integration of all routines resulted in CFIT. The version of CFIT used here was written in ANSI C and runs under UNIX on an IBM RISC 6000 computer. We have published a short software description of CFIT [28]. Fitting a spectrum with 400 data points with a 22-line mathematical model in a 1000-generation evolution period takes approximately 2.5 h calculation time.

#### 5.4. Configuration of the genetic algorithm

A way to optimize the configuration of the genetic algorithm is to make educated guesses based on the expertise of the practitioners. The traditional approach is to apply optimization based on full factorial designs. Since it is undesirable to create a meta-optimization problem, we pursued a practical combined approach which comprised the following steps.

(1) A synthetic spectrum  $S(\nu)$  is created by selecting random peak parameters within the search volume as defined in Table 2. Some noise is added to the synthetic spectrum. So  $S(\nu)$  has approximately the same complexity in search space as the

Table 3

Configuration of the GA, based on  $2^{6-2}$  fractional factorial design

Number_of_generations	500
Population_size <sup>a</sup>	200
Fitness_offset <sup>a</sup>	0
Fitness_mode	sigmoid
Scale_mode <sup>a</sup>	2.5
Fraction_elitism <sup>a</sup>	0.04
Recombination_mode	<b>B_MX</b>
Recomb_probability <sup>a</sup>	0.80
Number_of_breakpoints <sup>a</sup>	2
Mutation_mode	<b>B_M</b>
Mutation_probability	0.01
Encoding_resolution	9 bits (512 levels)
Decoding_mode	Gray_binary

<sup>a</sup> These parameters were optimized.

experimental IR spectra. However, the exact solution is known.

(2) A  $2^{6-2}$  fractional factorial design was executed.

(3) From the results of the factorial design, the best parameter setting was chosen, based on the lowest root mean square error (RMSE) between experimen-

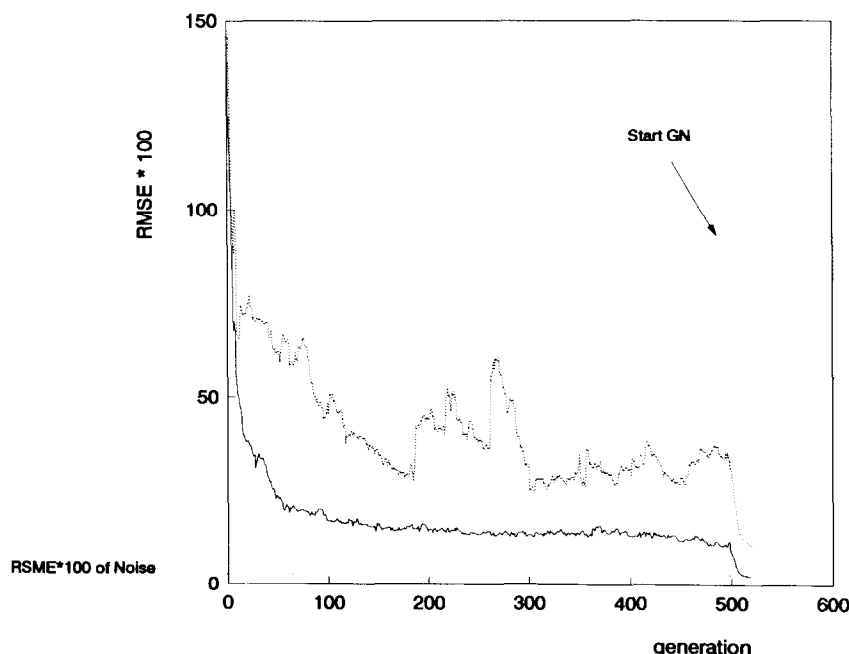


Fig. 10. The solid and dashed lines show the RMSE of the optimized and biased GA, respectively. After 500 generations the best solution found by the GA was post-hybridized by GN.

Table 4

RMSE of fitting of synthetic peak profile  $S(\nu)$  after 500 generations with a GA with optimal and biased parameter setting and subsequent iteration with SD

	RMSE $\times 100$ (GA)	RMSE $\times 100$ (SD)
Optimal GA	11.4	2.3
Biased GA	32.7	8.2

RMSE  $\times 100$  of noise is equal to 2.6.

tal and fitted spectrum after 500 generations. The settings are listed in Table 3.

In Fig. 10 two typical GA runs are depicted. The solid line represents the evolution of RMSE of fit for the best configuration of the GA as listed in Table 3. The dotted line shows RMSE evolution for configuration as in Table 3 but with number-of-breakpoints = 20 and population size = 50. After 500 generations we post-hybridized the ultimate GA solution with a steepest descent based curve fitting module with Gauss–Newton optimization, as described previously. Both GA solutions improved as a result of the sequential hybridization. However, the final

Table 5

RMSE values of the best solution in the population after  $n$  generations and the corresponding RMSE values of the subsequent GN fit

Generation number $n$	RMSE GA-best ( $\times 100$ )	RMSE GA-best $\rightarrow$ GN ( $\times 100$ )
0	255.6	26.48
1	193.3	29.54
5	154.7	20.51
10	110.7	13.72
25	90.29	4.850
35	67.75	4.850
50	60.86	14.73
75	52.14	4.850
100	48.54	4.850
150	41.14	3.525
250	35.66	3.525
500	18.74	3.525
1000	8.45	3.460

RMSE after steepest descent optimization of the poorly configured GA was considerably higher than the mean square error (MSE) of the noise added to  $S(\nu)$ . Also there was a considerable deviation of the

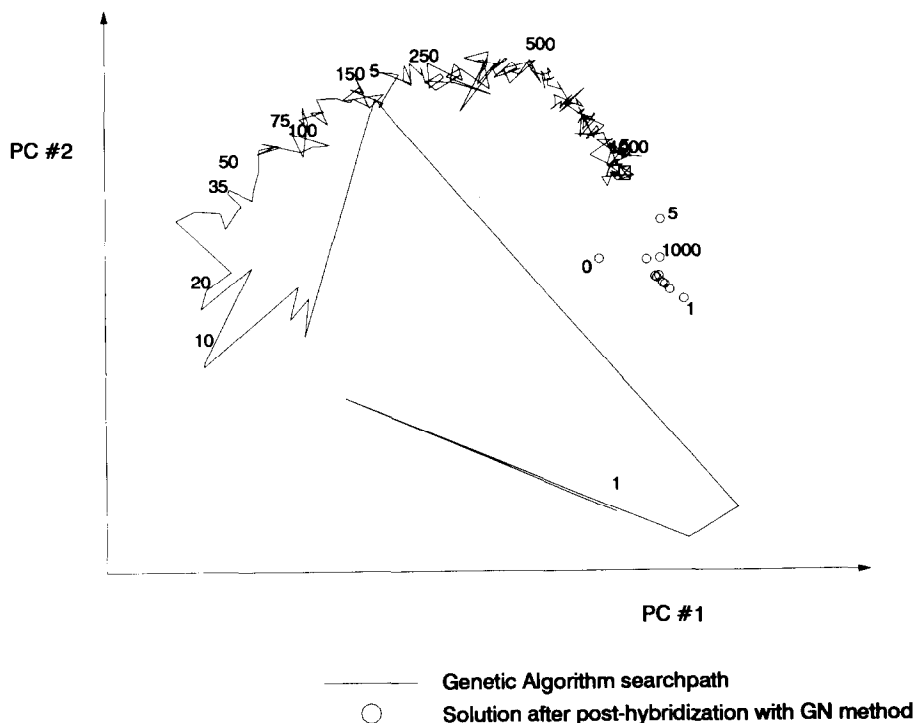


Fig. 11. Search trajectory of GA evolution in PCA score plot. The subsequent GN solutions consequently end in another region of the search space. The numbers refer to the RMSE values in Table 5.

Table 6  
Estimated peak areas in 14 repeated experiments

Approximate peak position	RMSE													s	area
	3.304	3.304	3.305	3.311	3.368	3.368	3.424	3.455	3.455	3.459	3.467	3.752	3.905		
1090	280.88	281.15	280.81	280.9	281.68	281.16	282.33	282.16	281.98	282.07	282.13	282.7	278.05	285.89	1.59
1043	12.46	12.42	12.47	12.45	12.38	12.45	12.3	12.37	12.39	12.38	12.37	12.3	12.88	11.44	0.28
1024	7.49	7.49	7.49	7.49	7.49	7.49	7.49	7.48	7.48	7.48	7.48	7.48	7.51	7.56	0.02
1021	14.31	14.32	14.31	14.3	14.34	14.32	14.34	14.41	14.4	14.4	14.4	14.41	14.17	13.84	0.14
1018	18.53	18.52	18.53	18.53	18.51	18.52	18.51	18.47	18.47	18.47	18.47	18.47	18.61	18.85	0.09
999	1.31	1.31	1.31	1.31	1.31	1.31	1.31	1.33	1.33	1.33	1.32	1.32	1.27	1.24	0.02
991	2.78	2.78	2.77	2.77	2.78	2.78	2.78	2.82	2.81	2.82	2.81	2.81	2.69	2.62	0.05
970	17.37	17.37	17.39	17.39	17.36	17.36	17.37	17.17	17.18	17.17	17.2	17.19	18.16	15.73	0.48
972	8.86	8.86	8.87	8.87	8.86	8.86	8.87	8.9	8.91	8.9	8.89	8.9	8.81	13.05	1.07
945	20.35	20.35	20.31	20.32	20.43	20.43	20.37	20.97	20.93	20.97	20.89	20.9	19.63	17.84	0.78
899	5.29	5.29	5.3	5.3	5.33	5.33	5.3	6.07	6.08	6.07	5.83	5.46	0.98	7.1	1.30
877	5.13	5.13	5.13	5.13	4.9	4.9	3.14	6.47	6.47	6.47	6.06	6.81	11.55	0.44	2.29
874	0.83	0.82	0.82	0.82	1.05	1.05	2.65	0.71	0.71	0.71	0.71	0.7	3.25	4.08	1.01
851	15.24	15.25	15.26	15.25	16.92	16.92	9.34	16.48	16.48	16.47	19.76	18.58	16.03	11.36	2.53
849	2.98	2.97	2.97	2.97	2.07	2.07	5.06	2.98	2.98	2.98	2	2.46	3.02	2.77	0.71
849	0.63	0.63	0.62	0.63	0	0	4.23	3.44	3.44	3.44	2.25	2.61	2.72	5.14	1.63
818	20.15	20.14	20.13	20.14	19.8	19.8	20.72	11.12	11.12	11.12	10.75	10.77	9.72	21.23	4.72
794	1.15	1.15	1.15	1.15	0.95	0.95	1.16	1	1	1	1	1	1.04	1.2	0.08
788	1.94	1.94	1.95	1.94	1.54	1.54	1.39	1.71	1.71	1.71	1.71	1.7	1.77	1.04	0.24
762	3.86	3.85	3.86	3.86	4.19	4.19	3.96	5.58	5.58	5.58	5.56	5.55	6.12	3.68	0.87
748	1.26	1.26	1.26	1.26	1.32	1.32	1.29	1.7	1.7	1.7	1.69	1.69	1.72	1.25	0.21
730	19.57	19.57	19.57	19.57	19.57	19.57	19.56	19.57	19.57	19.57	19.57	19.57	19.58	19.51	0.01

original peak parameters in the latter case. The configured GA reaches a perfect fit with a RMSE roughly equal to the MSE of the noise. The errors are summarized in Table 4.

### 5.5. Fitting of experimental profiles

With boundaries formulated around the initial estimates as written in Table 1 and 2, an evolutionary run of 1000 generations was exerted on the isotropic spectrum which had the lowest RMSE with respect to the mean spectrum. The parameters of the best fit in each population after respectively 1, 5, 10, 20, 35, 50, 75, 100, 150, 250, 500 and 1000 generations were used as initial estimates for a Gauss–Newton fit. Because several GN fits from GA estimates resulted in different converged final estimates we may conclude that many local optimal points were present in the search space (Table 5). The area in the solution space with the lowest RMSE values is not explored by the genetic algorithm since the RMSE value gradually decreases with the number of generations to a value which is clearly higher ( $\text{RMSE} \approx 8 \times 10^{-2}$ ) than the final estimates of the subsequent GN fits ( $\text{RMSE} \approx 3 \times 10^{-2}$ ). Both methods seem to optimize on a different ‘level’ of complexity. More evidence for this can be given by examination of the search trajectory. After each generation, the estimates of all 22 peak areas of the best solution in the population was stored. The resulting data matrix (dimension:  $22 \times 1000$ ) was subjected to principal component analysis.

In Fig. 11, a projection of the trajectory in search space is given in a PCA score plot. In this plot 75% of the original variation in peak area could be retained in two dimensions. The figure suggests that the genetic algorithm is capable to explore the search space (generation 0–40) and to exploit the search space (40–1000), without change in the parameter setting of the algorithm. During the evolutionary run, the various subsequent steepest descent like optimization procedures ended in local optima as shown in Table 5. These local optima are situated in a region of the search space area not explored by the genetic algorithm. If we assume that there is an ambiguity in search space, i.e. different combinations of peak parameters give an almost equally close fit to the experimental profile, the sets of peak parameters which give good fits are strongly correlated. Genetic

algorithms do not make use of correlations since they are only based on random mutations and cross-overs with preservation of the best solution. GN based procedures use correlation to optimize to a final estimate. Since the GN-RMSE values are correlated with the GA-RMSE values, we consider the search accuracy to be high for GA (in finding the region in which the absolute minimal RMSE is located) and the search precision to be high for GN (in finding the local minimal RMSE). By repeating the experiments 14 times (for different random initial populations) with 1000 generations and subsequent GN post-hybridization, it was observed that for most of the peak areas, the standard deviations of the fit are low (Table 6). Ambiguous fitting results may be expected in the  $877 \text{ cm}^{-1}$ ,  $851 \text{ cm}^{-1}$  and in the  $818 \text{ cm}^{-1}$  region. Although the RMSE values are low, the standard deviations of these peaks are unacceptably high.

## 6. Conclusions and outlook

We have shown that GAs feature a good search accuracy, although the search precision is poor. The opposite properties apply to the GN technique. The combination of both procedures combines good search accuracy and precision. Special attention has to be paid to the exchange from GA to GN. If the evolution period for the GA is too short, the resulting estimates for the GN method are not accurate enough to prevent ill-defined parameter values. If the evolution time is taken much longer than necessary, there is an appreciable waste of computation time. We have now fitted all spectra with the GA–GN procedure and quantitative relations between band parameters and structural features such as crystallinity and orientation were obtained. The interpretation of the results is beyond the scope of this paper and will soon be published elsewhere.

## Acknowledgements

We wish to acknowledge the following contributions to the work presented: ACR-A for financial support, D. Nijland for work on the artificial neural network for peak detection, C.J.M. van den Heuvel, who stimulated and advised based on his knowledge

of evaluation of the physical structure of synthetic yarns and C.B. Lucasius and H. Mannee for the development of the curve fitting programs.

## References

- [1] B.G.M. Vandeginste and L. de Galan, Critical evaluation of curve fitting in infrared spectrometry, *Anal. Chem.*, 47 (1975) 2124–2132.
- [2] J.A. Pierce, R.S. Jackson, K.W. Van Every, P.R. Griffiths and G. Hongjin, Combined deconvolution and curve fitting for quantitative analysis of unresolved spectral bands, *Anal. Chem.*, 62 (1990) 477–484.
- [3] R.J. Samuels, *Structural Polymer Properties*, Wiley-Interscience, New York, 1974.
- [4] A.P. de Weijer, L. Buydens, G. Kateman and H.M. Heuvel, Neural network used as a soft-modelling technique for quantitative description of the relation between physical structure and mechanical properties of poly(ethylene terephthalate) yarns, *Chemom. Intell. Lab. Syst.*, 16 (1992) 77–86.
- [5] H.M. Heuvel, L.J. Lucas, C.J.M. van den Heuvel and A.P. de Weijer, Experimental relations between physical structure and mechanical properties of a huge number of drawn PET yarns, *J. Appl. Polym. Sci.*, 45 (1992) 1649–1660.
- [6] H.M. Heuvel and R. Huisman, Infrared spectra of poly(ethylene terephthalate) yarns. Fitting of spectra, evaluations of parameters, and applications, *J. Appl. Polym. Sci.*, 30 (1985) 3069–3093.
- [7] C.J.M. van den Heuvel, H.M. Heuvel, W.A. Faassen, J. Veurink and L.J. Lucas, Molecular changes of PET yarns during stretching measured with rheo-optical infrared spectroscopy and other techniques, *J. Appl. Pol. Sci.*, 49 (1993) 925–934.
- [8] B. Jasse and J.L. Koenig, Orientational measurements in polymers using vibrational spectroscopy, *J. Macromol. Sci. Rev. Macromol. Chem.*, C17(1) (1979) 61–135.
- [9] K. Pearson, *Phil. Trans. A*, 186 (1895) 343.
- [10] P. Lapersonne, D.I. Bower and I.M. Ward, Molecular orientation and conformational changes due to uniaxial-planar deformation of poly(ethylene terephthalate) films, *Polymer*, 33 (1992) 1277–1283.
- [11] X. Yang, F. Long, D. Shen and R. Qian, Infra-red determination of conformers of poly(ethylene terephthalate), *Polymer Commun.*, 32 (1991) 125–128.
- [12] S. Lin and J.L. Koenig, *J. Polym. Sci. Polym. Phys. Ed.* 20 (1982) 2077–2088.
- [13] A.P. de Weijer, C.B. Lucasius, L. Buydens, G. Kateman, H.M. Heuvel and H. Mannee, Curve fitting using natural computation, *Anal. Chem.*, 66 (1994) 23–31.
- [14] A.P. de Weijer, L. Buydens, G. Kateman and H.M. Heuvel, Using genetic algorithms for an artificial neural network model inversion, *Chemom. Intell. Lab. Syst.*, 20 (1993) 45–55.
- [15] J. Zupan and J. Gasteiger, Neural networks: A new method for solving chemical problems or just a passing phase? (Review), *Anal. Chim. Acta*, 248 (1991) 1–30.
- [16] D. Rumelhart and J. McClelland, *Parallel Distributed Processing*, Vol. 1, MIT Press, Cambridge, MA, 1986.
- [17] I. Noda, Generalized two-dimensional correlation method applicable to infrared, Raman, and other types of spectroscopy, *Appl. Spectrosc.*, 47 (1993) 1329–1336.
- [18] F.E. Barton, Two-dimensional vibration spectroscopy: correlation of mid- and near-infrared regions, *Appl. Spectrosc.*, 46 (1992) 420–429.
- [19] T.R. Manley and D.A. Williams, *Polymer*, 8 (1967) 339.
- [20] W. Windig, D.E. Margevich and W.P. McKenna, A novel tool for two-dimensional (2D) correlation spectroscopy, *Chemom. Intell. Lab. Syst.*, in press.
- [21] N.R. Draper and H. Smith, *Applied Regression Analysis*, Wiley, New York, NY, 1967.
- [22] S. Röber and H.G. Zachmann, Determination of molecular orientation in poly(ethylene terephthalate) by means of  $^2\text{H}$  nuclear magnetic resonance, *Polymer* 33 (1992) 2061–2075.
- [23] D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, MA, 1989.
- [24] C.B. Lucasius and G. Kateman, Understanding and using genetic algorithms. Part 1: Concepts, properties and context, *Chemom. Intell. Lab. Syst.*, 19 (1993) 1–33.
- [25] C.B. Lucasius and G. Kateman, Understanding and using genetic algorithms. Part 2: Representation, configuration and hybridization, *Chemom. Intell. Lab. Syst.*, 25 (1994) 99–145.
- [26] D.B. Hibbert, Genetic algorithms in chemistry, *Chemom. Intell. Lab. Syst.*, 19 (1993) 277–293.
- [27] C.B. Lucasius and G. Kateman, GATES towards evolutionary large-scale optimization: a software-oriented approach to genetic algorithms. (a) Part 1: General perspective. (b) Part 2: Toolbox description, *Comp. Chem.*, 18 (1994) 137–156.
- [28] C.B. Lucasius, A.P. de Weijer, L. Buydens and G. Kateman, CFIT: a genetic algorithm for survival of the fitting. Software description, *Chemom. Intell. Lab. Syst.*, 19 (1993) 337–341.